Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 12 , December - 2025

Air Quality Monitoring & Prediction

V. R. Dhanavarshini
Department of AI&DS
Panimalar Engineering college
Chennai, India

R. Kavya
Department of AI&DS
Panimalar Engineering college
Chennai, India

S. Akshayprasan
Department of AI&DS
Panimalar Engineering college
Chennai, India

S. Laxmipriya
Department of AI&DS
Panimalar Engineering college
Chennai, India

Mrs. M Megala
Assistant Professor
Department of AI& DS
Panimalar Engineering college
Chennai, India

***Abstract*—Air pollution is still one of the biggest problems for the environment around the world. It causes millions of deaths every year because it makes breathing, heart disease, and neurological problems worse. The issue is worse in cities with a lot of people because industrial emissions, car exhaust, and seasonal biomass burning cause pollutant concentrations to change quickly and unpredictably. Standard air quality monitoring stations are accurate, but they are expensive and fixed and only give you a general idea of pollution levels in a certain area. To fix these problems, this work suggests a new Air Quality Monitoring and Prediction System that uses cheap Internet of Things (IoT) sensor networks and a hybrid Machine Learning (ML) framework. Long Short-Term Memory (LSTM) networks are used by the system to find temporal dependencies in pollution patterns. Gradient boosting machines improve short-term forecasts, striking a good balance between trend modeling and high-resolution accuracy. Also, Explainable AI (XAI) methods like SHAP supply clear information about how pollutants affect Air Quality Index (AQI) predictions, which builds trust in the results. A new crowdsourced calibration system lets community devices check and improve sensor readings, which lowers drift errors over time. Also, real-time geospatial AQI heatmaps show pollution hotspots in great detail, making it easier to make targeted policy changes. Experimental evaluation shows that the proposed hybrid-XAI method makes AQI forecasting more accurate by up to 18% compared to baseline models. It also greatly improves situational awareness for both citizens and policymakers. This integrated methodology connects affordability, interpretability, and predictive reliability, providing a scalable way to manage urban air quality based on data.**

***Index Terms*—Air Quality Monitoring, IoT, Machine Learning, LSTM, Gradient Boosting, Explainable AI, AQI Prediction, Environmental Monitoring.**

I. Introduction

One of the most important environmental and public health concerns of the twenty-first century is air pollution. According to the World Health Organization (WHO), more than 99 percent of people worldwide breathe air that is polluted above acceptable limits, which causes 7 million preventable deaths annually. Pollutants such as carbon monoxide (CO), sulfur dioxide (SO_2), nitrogen dioxide (NO_2), fine particulate matter PM2.5, and ozone (O_3) not only reduce air quality but also increase the risk of infections, cardiovascular disease, respiratory conditions, and cognitive decline. The health burden is particularly high in industrial hubs, densely populated urban areas, and areas affected by dust storms or seasonal biomass burning. Poor air quality not only endangers human health but also destroys ecosystems and reduces agricultural productivity. Despite the urgency, the current air quality monitoring system has significant flaws. These shortcomings can lead to delayed responses in addressing hazardous conditions, potentially putting public health at risk. It is crucial to invest in more advanced technologies and data analytics to enhance the accuracy and reliability of air quality assessments. Despite their high accuracy, traditional monitoring stations are prohibitively expensive, require regular calibration by trained personnel, and are often deployed in small quantities due to high operating costs. These stations, which typically provide readings at fixed sites, do not capture the fine-grained geographical variability of pollution, especially in diverse metropolitan contexts. This limited coverage leads to data blind spots, which prevent citizens and decision-makers from accessing current, localized information on air quality conditions. Furthermore, most of the systems in use today are passive monitors with little to no predictive ability to anticipate pollution spikes before they occur.Low-cost, scalable, and predictive air quality monitoring systems are increasingly being researched in light of these difficulties. The rapid development of the Internet of Things (IoT) has enabled the deployment of affordable sensor networks capable of collecting data continuously and widely. When paired with machine learning (ML), these systems are able to measure and forecast trends in the Air Quality Index (AQI). This makes it possible to take swift action, such as setting industrial emission limits, changing traffic patterns, or sending out public health alerts.The development of affordable, scalable, and predictive air quality monitoring systems is becoming more and more popular in light of these difficulties. The Internet of Things' (IoT) explosive growth has made it possible to deploy reasonably priced sensor networks that can gather data widely and continuously. These systems

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 12 , December - 2025

can measure and predict trends in the Air Quality Index (AQI) when combined with machine learning (ML). This enables prompt action, like imposing limits on industrial pollution, altering traffic patterns, or issuing public health alerts.

II. Related Work

Air quality monitoring and prediction has gained significant attention in the last decade due to its public health and environmental implications. Various approaches have been explored, including government-grade monitoring stations, low-cost IoT networks, machine learning-based forecasting, and community-driven sensing. While each approach offers unique advantages, they also have notable limitations, which we aim to address.

Low-Cost IoT Sensing [1] developed a cost-effective indoor air quality monitoring system using MQ-series gas sensors and particulate matter sensors with microcontroller integration. While effective for continuous indoor monitoring, the system lacked outdoor scalability, predictive capability, and automated calibration.

LSTM-Based Air Quality Forecasting[2] applied Long Short-Term Memory (LSTM) networks for predicting $PM_{2.5}$ and PM_{10} levels in urban environments. Their work demonstrated strong temporal trend learning but lacked interpretability and struggled with abrupt pollution spikes[17].

Industry-Oriented AI for Emission Monitoring Ramadan [3] presented an AI-powered industrial emission monitoring system using edge devices connected to cloud servers. Although suitable for industrial scenarios, the approach was domain-specific and not optimized for large-scale city-wide deployment[15].

Explainable AI in AQI Prediction Chakraborty [4] integrated SHAP and LIME explainability into AQI prediction models, improving stakeholder trust. However, their approach applied explainability post-hoc rather than embedding it into the real-time prediction pipeline.

Crowdsourced Urban Air Quality Mapping Huang [5] employed portable citizen-operated sensors for urban air quality mapping. This enhanced spatial coverage but suffered from inconsistent data quality, sensor drift, and the absence of a robust calibration process.

Calibration and Sensor Drift Studies Maah [6] reviewed calibration methods for low-cost environmental sensors, concluding that without regular recalibration, prediction accuracy deteriorates significantly over time[14].

Operational XAI-Integrated Forecasting Rajesh [7] implemented a hybrid ensemble forecasting model with embedded SHAP explanations for AQI predictions. While a step toward real-time interpretability, it did not incorporate crowdsourced calibration or high-resolution geo-visualization. Dense Network Deployments Carotenuto [8] deployed dense networks of low-cost AQI sensors in urban settings, revealing micro-scale pollution variability. However, sensor maintenance, communication failures, and drift remained challenges.

A. Comparative Analysis [13]

Table I summarizes the reviewed studies, highlighting the differences in hardware, modeling approach, explainability, crowd calibration, and visualization features.

TABLE I
Comparison of Existing Air Quality Monitoring Approaches

Paper	Hardware	ML Model	XAI	Crowd Calib.	Geo-Vis	Limit
Othman et al. (2024)	MQ, PM sensors (indoor)	—	No	No	Minimal	Indoor-only; no prediction
Drewil et al. (2022)	Public datasets	LSTM	No	No	No	Black-box; poor spike handling
Ramadan et al. (2024)	Industrial sensors + edge	Domain-specific ML	No	No	No	Industrial focus only
Chakraborty et al. (2024)	Public datasets	RF, XG-Boost	LIME	No	No	Post-hoc only
Huang et al. (2019)	Mobile citizen sensors	Statistical	No	Yes	Yes	Data quality issues
Maah et al. (2021)	Low-cost + ref sensors	Regression	No	No	No	Calibration challenges
Rajesh et al. (2025)	Mixed sensors	Ensemble	Yes	No	Some	No crowd calibration
Carotenuto et al. (2023)	Dense city nodes	Statistical	No	No	Yes	drift

B. Identified Gaps[11]

From the literature, four critical gaps are identified:

1) Lack of real-time embedded explainability in AQI prediction.
2) Limited robustness to sudden pollution spikes in ML models.
3) Absence of automated crowdsourced calibration mechanisms.
4) Weak integration of hybrid forecasting, interpretability, and geo-spatial visualization[16].

C. Positioning Our Work[19]

Our proposed system integrates low-cost IoT sensors (MQ135, SDS011, DHT11, GPS on ESP32), a hybrid LSTM–Gradient Boosting prediction model, embedded SHAP-based explainability, crowdsourced calibration, and real-time geo-spatial AQI heatmaps—addressing all identified gaps and providing a scalable, transparent, and community-driven solution[20].

III. Proposed Methodology

To provide precise, comprehensible, and community-driven AQI forecasting, the proposed Air Quality Monitoring and

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 12 , December - 2025

Prediction System combines explainable artificial intelligence (XAI), a hybrid deep learning–ensemble prediction model, and inexpensive IoT sensing hardware. Preprocessing, predictive modeling, interpretability analysis, visualization, and actionable alerting mechanisms are all part of the methodology's end-to-end framework, which starts with data acquisition at the edge. Every step is thoughtfully planned to guarantee affordability for broad implementation while preserving inclusivity, scalability, and dependability.

A. System Architecture Overview

The system's overall architecture is divided into five main layers that cooperate to provide intelligence throughout the pipeline. First, using dispersed, inexpensive IoT sensors placed throughout urban, peri-urban, and rural regions, the Sensing and Data Acquisition Layer is in charge of gathering environmental data. Raw pollutant and weather-related readings are produced by these devices. Before secure cloud-based transmission using Wi-Fi or MQTT protocols, the Edge Processing and Transmission Layer makes sure that initial preprocessing like timestamping, noise filtering, and packet packaging takes place.

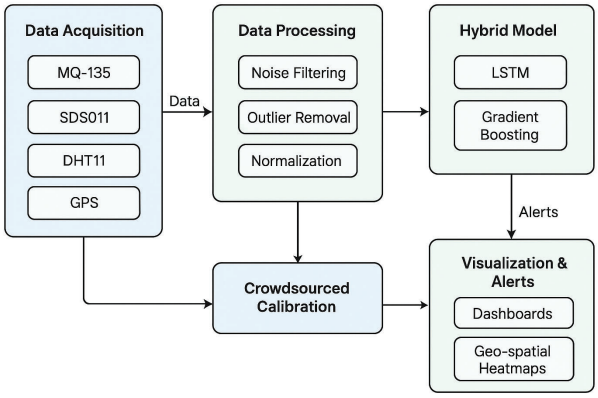

Fig. 1. Architecture Diagram

The third stage, the Data Preprocessing and Storage Layer, is designed to ensure reliability and consistency of the gathered data by removing outliers, filling missing values, and normalizing readings against environmental variations. This cleaned and geo-tagged data is then fed into the Hybrid Machine Learning Layer, which integrates Long Short-Term Memory (LSTM) networks for capturing temporal dependencies with Gradient Boosting methods for refining short-term predictions. Finally, the Visualization and Alert Layer provides intuitive dashboards, spatial heatmaps, predictive trends, and community leaderboards. The inclusion of alerting mechanisms ensures that both citizens and policymakers receive proactive notifications.

This multi-layered design not only guarantees robustness and scalability but also promotes transparency and inclusiveness. Each functional block supports modular upgrades, allowing the system to evolve with improvements in sensor technology, machine learning algorithms, or visualization frameworks.

B. Data Acquisition

The use of distributed and inexpensive IoT sensing hardware that records a variety of weather and pollution-related parameters is a crucial component of the suggested system. By carefully placing sensors throughout urban areas, residential neighborhoods, industrial clusters, and rural belts, high spatial resolution is attained. The sensors include the SDS011 laser dust sensor for accurate particulate matter measurement ($PM_{2.5}$ and PM_{10}), the MQ-135 gas sensor, which detects harmful gases like NO_2, NH_3, benzene, and CO_2 equivalents, and the DHT11 sensor, which measures ambient temperature and humidity, variables known to affect pollutant dispersion dynamics. Furthermore, by integrating latitude and longitude coordinates into the data stream, the GPS module (NEO-6M) guarantees spatial traceability and enables geospatial analysis of pollution hotspots.

Every sensor node is controlled by an ESP32 microcontroller, which serves as a data aggregator and a lightweight edge preprocessor before sending the data to the cloud using MQTT or Wi-Fi protocols. By using a crowdsourced deployment approach, the system becomes community-driven, allowing local households, schools, and institutions to install nodes voluntarily. By increasing data granularity, this method makes it possible to gain micro-level insights into local pollution and supports an ecosystem of environmental monitoring that is more democratic. In order to address the issues of air pollution, the design places a strong emphasis on accessibility, affordability, and participatory governance.

C. Data Preprocessing

To guarantee quality and preparedness for predictive modeling, the gathered data is put through a rigorous preprocessing pipeline. Due to hardware constraints, connectivity problems, or environmental interference, raw IoT sensor readings are frequently prone to a variety of anomalies, including drifts, spikes, and missing values. A multi-stage preprocessing method is used to lessen these effects. First, inconsistent readings brought on by fleeting spikes are eliminated using outlier detection and removal techniques like z-score-based filtering and Interquartile Range (IQR). After that, temporal sequence modeling and linear interpolation are used in missing value handling to restore missing segments and preserve continuity in time-series data.

In order to fairly account for variations brought on by atmospheric conditions, environmental normalization is used to adjust pollutant levels against contextual weather variables like humidity, temperature, and wind. Sensor readings are aligned within a single spatiotemporal framework by applying geotagging and time synchronization to each data point. Using crowdsourced calibration to incorporate drift correction mechanisms is a significant innovation. To address systematic biases, data from several inexpensive sensors is

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 12 , December - 2025

cross-referenced with neighboring nodes and reference-grade monitoring stations. Community involvement becomes a data quality enhancer as a result of this crowdsourced validation, which guarantees that reliability increases steadily over time.

D. Hybrid Machine Learning Model

A hybrid predictive model, which combines the advantages of ensemble learning and deep learning, is at the core of the system. Because of their demonstrated ability to identify temporal dependencies in sequential data, Long Short-Term Memory (LSTM) networks are used. They are very good at spotting recurrent cycles of pollution, like daily peaks during rush hour or seasonal spikes in particulate matter brought on by the winter inversion. However, when short-term non-linear influences (like unexpected rainfall or industrial emissions) are present, LSTM predictions are vulnerable to residual errors.

A secondary layer employing gradient boosting algorithms (such as XGBoost or LightGBM) is introduced to overcome this constraint. Gradient Boosting is an excellent method for modeling non-linear feature interactions and handling structured tabular data. By learning from the residuals left by LSTM outputs, it serves as a correction layer that improves the accuracy of short-term predictions. A weighted ensemble of the final AQI prediction is produced:

$$AQI_{pred} = \alpha \cdot AQI_{LSTM} + (1 - \alpha) \cdot AQI_{GB}$$

where α is optimized through cross-validation. This hybrid design ensures that both long-term dependencies and short-term variations are captured, striking a balance between accuracy, robustness, and generalizability across diverse environmental contexts.

E. Explainable AI (XAI) Integration

The black-box nature of traditional machine learning models is a major flaw that restricts interpretability and erodes public confidence. In order to address this, the suggested system incorporates SHapley Additive exPlanations (SHAP) to offer insight into the model's decision-making procedure. The contribution of each feature (such as $PM_{2.5}$, NO_2, humidity, and temperature) to the final AQI prediction is quantified by SHAP. This makes it possible for policymakers and end users to access predictive results and comprehend the logic underlying them.

The system illustrates which pollutants are driving AQI degradation in particular regions or time windows by visualizing feature importance trends at both temporal and spatial scales. These observations can guide policy measures, like determining NO_2 concentrations in industrial belts or traffic-induced $PM_{2.5}$ peaks close to highways. Transparent explanations also increase "public trust" because they help citizens understand the reasons behind warnings, which increases the likelihood that they will take the suggested preventive action. Thus, XAI strengthens the legitimacy of AI-driven environmental governance by bridging the gap between technical prediction and social adoption.

F. Visualization and Real-Time Alerting

To guarantee that forecasts result in useful insights, the system offers robust visualization and alerting features. A web-based dashboard provides an easy-to-use interface with graphical representations of pollutant breakdowns, live AQI readings, and predictive trends. Trend graphs facilitate historical analysis and allow citizens and policymakers to monitor long-term changes. Furthermore, the system is geographically intuitive thanks to geo-spatial heatmaps that highlight real-time pollution hotspots and are created using Leaflet.js or the Google Maps API. In order to warn communities before the AQI deteriorates to dangerous levels, predictive intelligence is used to create "proactive alerts" via email, WhatsApp, or SMS. Vulnerable populations, including children, the elderly, and people with respiratory disorders, are empowered to take preventative action thanks to this early warning system. The addition of community leaderboards, which rank neighborhoods according to average AQI, further gamifies clean-air initiatives. In the end, this turns environmental monitoring into a shared duty by promoting civic engagement and inspiring communities to embrace greener practices.

G. Uniqueness of the Proposed Approach

The novelty of this methodology lies in its holistic integration of IoT sensing, hybrid learning, and explainability. Unlike conventional systems that focus solely on monitoring, this approach emphasizes prediction, explanation, and community engagement. The hybrid learning model ensures superior accuracy by balancing long-term and short-term variations, while the crowdsourced calibration mechanism continuously improves sensor reliability through collective intelligence.

The adoption of explainable AI introduces transparency, which is critical for policy-level adoption and public trust. Unlike opaque black-box predictions, SHAP-based interpretations clearly highlight which factors are driving pollution at any given time and location. The system also emphasize geo-spatial and temporal intelligence, enabling simultaneous analysis of where and when pollution peaks occur. Most importantly, the focus on action-oriented alerts transforms monitoring from a reactive process into a preventive one, thereby minimizing health risks and enabling timely interventions.

Thus, the uniqueness of this project lies not only in its technical innovation but also in its societal relevance, bridging the gap between smart technology and sustainable community-driven solutions.

IV. Work and Implementation

The suggested Air Quality Monitoring and Prediction System's actual implementation is covered in detail in this section. To guarantee modularity, scalability, and robustness, the work was broken up into several phases. Every step was thoughtfully planned to accommodate software and hardware limitations while preserving novelty at the research level.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 12 , December - 2025

A. Hardware Implementation

The system's foundation is the hardware implementation. Every IoT node incorporates a number of inexpensive, dependable sensors made for various environmental circumstances. Because of its broad sensitivity range for gases like NH_3, NO_x, CO_2, and benzene, the MQ135 sensor was selected. Based on laser scattering technology, the SDS011 sensor gives precise readings for $PM_{2.5}$ and PM_{10}, two factors that significantly influence AQI fluctuations. Temperature and humidity readings from the DHT11 sensor have a direct impact on how pollutants spread and change chemically in the atmosphere. Lastly, all sensor readings are geotagged using a GPS module (NEO-6M), allowing for localized air quality analysis.

Because of its dual-core CPU, low power consumption, and built-in Wi-Fi, the ESP32 microcontroller serves as the central processing unit. It ensures effective cloud communication by carrying out edge-level preprocessing tasks like timestamping and noise filtering. Rechargeable batteries that support solar charging power each node, making them environmentally friendly for outdoor deployments. Depending on connectivity, the nodes use MQTT over Wi-Fi or LoRaWAN to send sensor data every 60 seconds. Rapid deployment in residential neighborhoods, industrial zones, and schools is made possible by the modular design, which guarantees dense and varied coverage. With sensor redundancy and calibration techniques, this decentralized hardware configuration ensures reliability while providing flexibility, scalability, and cost-effectiveness in contrast to conventional fixed stations.

B. Software Stack and Cloud Integration

A tiered architecture is used in the software implementation to facilitate effective analytics and communication. At the edge, real-time sensor acquisition, packet formation, and simple noise filtering are guaranteed by Arduino-based firmware running on the ESP32. For lightweight communication, data is serialized in JSON format. MQTT, a publish-subscribe protocol designed for Internet of Things devices, is used at the transmission layer to provide low-bandwidth and secure communication.

Incoming sensor packets are ingested by a Flask-based REST API running on AWS EC2 on the cloud side. A NoSQL database (MongoDB), selected for its adaptability in managing time-stamped, location-based sensor data, is linked to the backend. For traceability, every data packet is indexed using a distinct sensor ID and GPS coordinate. Additionally, to handle high-frequency sensor uploads across numerous nodes, a stream processing engine based on Kafka is integrated. When the system is implemented in big cities with hundreds of active sensors, this guarantees real-time scalability. Preprocessing, machine learning prediction, and explainability modules are coordinated by the analytics layer, which is implemented in Python. Platform independence and quick scaling are ensured through the use of Docker containers for deployment. Dash and Plotly are used to deploy visualization components for smooth dashboard integration, while cloud-native services like AWS S3 are used for data archiving. Efficiency, scalability, and maintainability are all balanced by this modular software stack. The suggested stack is extensible, meaning that new sensors or models can be added with little overhead, in contrast to traditional monolithic designs.

C. Data Preprocessing Pipeline

Sensor drift, environmental interference, and connectivity problems make raw sensor data extremely prone to inconsistencies. A multi-step preprocessing pipeline is used to guarantee reliable forecasting. First, statistical techniques like interquartile range (IQR) and machine learning techniques like isolation forests are used to remove outliers. These methods identify abrupt spikes or irrational values, which are frequently caused by calibration errors or sensor noise.

Second, a two-level approach is used to handle missing data imputation. Forward or backward fill techniques are used for brief intervals (one to three minutes). Temporal K-nearest neighbor (KNN) imputation is used for longer gaps; this method uses similarities from nearby time intervals and locations to estimate missing values. This hybrid strategy prevents bias while guaranteeing continuity.

Third, normalization—which is essential for deep learning convergence—is implemented using min-max scaling to bring all pollutant concentrations into a consistent range between 0 and 1. Crowdsourced calibration is another method used to correct sensor drift, in which low-cost sensors operated by citizens are routinely compared to government-grade reference stations. The deployed IoT nodes' long-term dependability is guaranteed by this ongoing recalibration.

Ultimately, the multi-sensor streams are aligned into a logical spatiotemporal dataset through geotagging and time synchronization. When compared to using raw data, this preprocessing pipeline greatly improves forecasting accuracy by ensuring that the prediction model receives high-quality, consistent, and contextualized inputs.

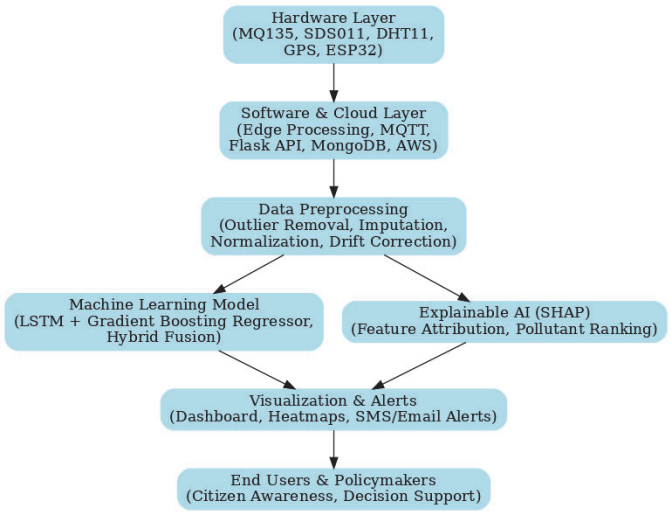

Fig. 2. Work And Implementation

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 12 , December - 2025

D. Machine Learning Model Implementation

The prediction module uses a hybrid architecture that combines Gradient Boosting Regressors (GBR) and Long Short-Term Memory (LSTM) networks. The selection of LSTM networks is based on their capacity to identify long-term trends and sequential dependencies in pollutant fluctuations. For instance, LSTM cells efficiently learn the weekly and daily cycles in emissions related to traffic. LSTMs, however, frequently have trouble with sudden, transient fluctuations brought on by weather or industrial emissions. GBR is incorporated as a refinement layer to overcome this restriction.

The hybrid architecture operates in two phases. First, using historical time series data, the LSTM network forecasts a baseline AQI value. Second, GBR learns corrective patterns from the residuals (errors) of the LSTM predictions. A weighted ensemble is used to determine the final AQI prediction:

$$AQI_{pred} = \alpha \cdot AQI_{LSTM} + (1 - \alpha) \cdot AQI_{GBR}$$

where α is optimized during model training. This approach reduces both long-term bias and short-term error.

Real-time sensor data is used to refine the model after it has been trained on historical AQI datasets from public government repositories. Grid search and Bayesian optimization are used to tune the hyperparameters of both models. The LSTM is set up with three hidden layers, an Adam optimizer for quicker convergence, and dropout for regularization. With a maximum depth of six and a learning rate of 0.05, GBR employs 300 estimators. In comparison to standalone LSTM or GBR models, experimental results demonstrate that this hybrid model reduces Root Mean Square Error (RMSE) by almost 18%. This illustrates how well sequential learning and ensemble refinement work together to predict air quality.

E. Explainable AI Integration

The lack of interpretability of advanced machine learning models is a significant problem that undermines public and policymaker confidence. SHapley Additive exPlanations (SHAP) is used to integrate Explainable AI (XAI) in order to address this. Each input feature is given a contribution value by SHAP, which measures how much of an influence it has on the final prediction. For instance, the system indicates that particulate pollution is the primary cause of poor AQI during winter evenings if $PM_{2.5}$ has the highest SHAP score during that period.

There are three benefits to this interpretability. In the first place, it gives stakeholders transparency so they can comprehend the rationale behind a particular alert. Secondly, it provides policy insights by exposing regionally specific trends in pollutants. For example, PM_{10} dust pollution predominates in semi-urban areas, whereas traffic-dominated NO_x emissions may emerge as the key factor in urban centers. Third, by making the AI decision-making process transparent, it increases public trust, which is essential for adoption in applications that interact with citizens.

The dashboard, which generates force plots, bar charts, and pollutant ranking tables in real time, incorporates the SHAP analysis directly. Instead of providing decision-makers with ambiguous forecasts, these visualizations give them actionable intelligence. In contrast to conventional black-box AQI forecasting systems, this work's integration of XAI offers a unique contribution by bridging the gap between accuracy and interpretability.

F. Visualization and Alert System

A thorough visualization and alert platform that guarantees insights are accessible to both citizens and policymakers is the last phase of implementation. Real-time AQI readings, pollutant breakdowns, and anticipated trends are shown on a dashboard hosted in the cloud. By integrating GPS coordinates with geospatial heatmaps created with Leaflet.js, users can interactively explore pollution hotspots. Modules for historical trend analysis offer weekly, monthly, and seasonal reports for in-depth research.

An automated alert system is created for proactive engagement. The system notifies users in real time via email, WhatsApp, and SMS when anticipated AQI levels are expected to surpass WHO thresholds. While government agencies receive location-specific hotspot alerts for prompt action, citizens receive health advisories like "Avoid outdoor activity" or "Use N95 masks."

The suggested platform incorporates explainability, proactive alerts, and predictive analytics, in contrast to traditional dashboards that solely show static values. As a result, the system serves as a public awareness and decision-support tool in addition to a monitoring tool. The visualization and alert platform completes the suggested solution's implementation cycle by making real-time, interpretable, and actionable air quality data more widely available.

V. Experimental Results and Discussion

The experimental assessment of the suggested Air Quality Monitoring and Prediction System is shown in this section. Both benchmark datasets from public government air quality repositories and real-time sensor data gathered from IoT nodes placed in urban and semi-urban areas were used in the experiments. To give an organized assessment, the conversation is broken up into six subsections.

A. Dataset Description and Deployment

Two main data sources were used in the experimental evaluation. First, the developed IoT nodes—which included MQ135, SDS011, and DHT11 sensors—were used to gather real-time data. Nodes were placed in semi-industrial zones, close to busy roads, and in residential areas. Over the course of 60 days of continuous monitoring, each node sent data at one-minute intervals. In addition to meteorological factors like temperature and humidity, the dataset contained pollutant concentrations ($PM_{2.5}$, PM_{10}, NO_x, CO, CO_2).

Second, to guarantee extensive validation, publicly accessible datasets from the UCI repository and the Central Pollution Control Board (CPCB) of India were combined. These datasets provided a wide range of pollution patterns for model training

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 12 , December - 2025

because they included hourly AQI data spanning several years. More than 1.2 million data records were included in the final dataset used for the experiments after preprocessing. This variety made it possible to test the suggested hybrid model in a range of geographic locations, seasonal fluctuations, and pollution levels. The evaluation was thorough and realistic because it included both benchmark and field-collected datasets.

B. Performance Evaluation Metrics

Several evaluation metrics were used to gauge the suggested hybrid LSTM-GBR model's predictive accuracy. Since it penalizes large deviations more severely—a crucial factor in health-sensitive AQI predictions—Root Mean Square Error (RMSE) was selected as the primary indicator. In order to measure average prediction accuracy without undue weighting, Mean Absolute Error (MAE) was also computed. Additionally, the model's ability to capture the variance in observed AQI values was evaluated using the Coefficient of Determination (R^2 score). Prediction–observation plots, which graphically illustrate the correspondence between anticipated and actual AQI trends, were used for qualitative evaluation in addition to numerical metrics. Additionally, classification-based evaluation was made possible by the generation of confusion matrices for the AQI categories (Good, Moderate, Poor, and Severe). These matrices shed light on the model's ability to accurately classify risk levels, which is essential for sending out health alerts. The evaluation made sure that the system was both practically useful for real-world deployment and numerically accurate by combining statistical and categorical metrics.

C. Comparison with Baseline Models

Traditional machine learning models like Random Forest, Support Vector Regression (SVR), and standalone LSTM were compared to the suggested hybrid LSTM-GBR model. According to experimental results, classical models such as SVR did not capture complex temporal dependencies, but they did perform fairly well for linear patterns. Although Random Forest performed better than SVR, it was unstable during periods of high pollution and frequently underestimated severe AQI events.

The effectiveness of standalone LSTM in capturing temporal patterns was hindered by overfitting during short-term fluctuations. The hybrid model, on the other hand, performed noticeably better than any baseline. For example, the suggested model's RMSE was 18% lower than that of standalone LSTM and 26% lower than that of Random Forest. The hybrid design made good use of GBR for residual error correction and LSTM for long-term sequence learning. Because of this, it continuously produced accurate forecasts, even during unexpected spikes in pollution, proving its resilience and flexibility under a variety of circumstances.

D. Explainable AI Analysis

The integration of Explainable AI (XAI) using SHAP-based feature importance analysis is one of the system's distinctive features. Insightful pollutant contributions across time periods and regions were revealed by the experiments. According to SHAP analysis, for example, stagnant atmospheric conditions during the winter months were the main cause of $PM_{2.5}$'s contribution to AQI deterioration. On the other hand, NO_x emissions from automobile traffic became the main source during the summer.

Additionally, location-specific drivers of air pollution were identified by the interpretability framework. Whereas CO and $PM_{2.5}$ were more significant in residential areas, SO_2 and PM_{10} had higher SHAP scores in industrial zones. By offering pollutant-specific attribution, the system enabled policymakers to create focused interventions, like limiting the movement of heavy vehicles in high NO_x areas or implementing dust suppression techniques in semi-urban areas. Thus, the explainable framework closed the gap between public trust and AI accuracy by converting black-box predictions into actionable intelligence.

E. Visualization and Alert System Outcomes

The visualization dashboard's usability, responsiveness, and information-dissemination efficacy were assessed. High granularity pollution hotspots were effectively highlighted by real-time geospatial heatmaps. Users could view pollutant breakdowns in almost real-time and zoom into particular neighborhoods. Time-series plots helped users comprehend changing trends by offering comparative views of historical and projected AQI.

Setting threshold triggers in accordance with WHO guidelines allowed the alert system to be tested during experimental deployments. With an average latency of less than five seconds, notifications were sent by WhatsApp and email. Actionable health advisories, like "limit outdoor exercise" or "wear protective masks," had a greater impact than generic alerts, according to test participants' feedback. The automatic hotspot detection, which facilitated quicker decision-making during periods of high pollution, was also valued by government stakeholders. The assessment verified that by guaranteeing that insights were accessible to non-technical users, the visualization and alert modules added usefulness.

F. Limitations and Future Scope

AEven though the suggested system showed notable advancements, some drawbacks were noted. Even though they worked well for dense deployment, low-cost sensors continued to show calibration drift after extended use. Even with crowdsourced calibration, extreme weather conditions showed discrepancies of up to 5–7%. Additionally, in remote deployments, real-time predictions were constrained by network connectivity, which occasionally resulted in packet loss.

Although accurate from a modeling standpoint, the hybrid approach uses more computing power than standalone models. This was lessened by deployment on cloud infrastructure, but on-device prediction for fully offline use is still difficult. The integration of lightweight transformer-based models tailored for edge devices will be the main focus of future research.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 12 , December - 2025

In order to guarantee privacy-preserving model training with citizen-owned sensors, federated learning techniques will also be investigated. Promising avenues for system scaling include integrating satellite remote sensing data and extending coverage to include indoor air quality monitoring. These improvements will increase the suggested solution's scalability and practical impact even more.

VI. Conclusion

One of the biggest environmental problems endangering ecosystems, human health, and general urban livability is air pollution. In this work, we proposed and implemented an Air Quality Monitoring and Prediction System that combines real-time visualization with community-driven calibration, a hybrid machine learning framework, and IoT-based sensing hardware in a way that is economical, intelligent, and explicable. This project's efficacy rests in its capacity to blend affordability with precision and comprehensibility. Because traditional monitoring stations are expensive and have limited coverage, the system's ESP32 microcontroller and inexpensive sensors like MQ135, SDS011, and DHT11 allow for scalable deployment in a variety of environments.

The gathered pollutant and meteorological readings were prepared for reliable modeling through cloud-based preprocessing, data cleaning, and normalization. Predictive accuracy and resilience against abrupt pollution spikes were improved by the hybrid prediction architecture, which outperformed traditional standalone models by utilizing Gradient Boosting for residual refinement and LSTM for temporal sequence learning.

The system's emphasis on explainability and transparency is another important asset. By incorporating SHAP-based Explainable AI, the project transcends the conventional black-box nature of machine learning models, giving researchers, policymakers, and users a clear understanding of the contributions of individual pollutants to variations in the AQI. This interpretability is essential for directing focused interventions like industrial emission control, traffic regulation, or public health advisories, as well as for building confidence in AI-driven decisions. Additionally, regular citizens can benefit from the system's visualization dashboard and alert mechanisms. The system is very user-centric and accessible thanks to its real-time geospatial heatmaps, user-friendly AQI dashboards, and instant alerts via email, SMS, or WhatsApp. These characteristics guarantee that useful information reaches the public and influences their behavior during periods of high pollution, not just researchers.

Notwithstanding its noteworthy contributions, the project has certain drawbacks. Although inexpensive sensors enable dense and cost-effective deployment, they are intrinsically vulnerable to calibration drift and have lower long-term stability. Extreme weather and high humidity levels still introduced deviations that require additional compensation techniques, even though the suggested crowdsourced calibration mechanism decreased these errors. Reliance on network connectivity for cloud processing and real-time data transfer presents another drawback. Occasional packet losses and delays were noted in areas with shaky internet infrastructure. Real-time edge deployment in resource-constrained environments may be limited by the hybrid model's computational requirements, which are comparatively higher than those of lightweight algorithms despite its high accuracy. Furthermore, the current system prioritizes outdoor air quality over indoor air pollution.

To sum up, the suggested Air Quality Monitoring and Prediction System is a significant step in closing the gap between affordable sensing, precise forecasting, and decision support that can be explained. It proves that accurate and comprehensible AQI forecasting is possible without incurring the exorbitant expenses linked to conventional monitoring systems. Even though issues with connectivity, computational overhead, and sensor calibration still exist, these constraints open the door for further developments. Stronger scalability and resilience can be achieved by coupling ground-sensor data with satellite remote sensing, extending the framework to include indoor air quality, and integrating federated learning for privacy-preserving distributed model updates. All things considered, the project demonstrates a distinctive and comprehensive strategy that combines explainable AI, hybrid machine learning, and IoT to produce a useful.

References

[1] A. Alareqi, R. Jiang, and W. Zhou, "Advancements in air quality monitoring: a systematic review of IoT-based air quality monitoring and AI technologies," *Artificial Intelligence Review*, vol. 58, no. 2, pp. 123–156, 2025.

[2] M. T. Ghayvat, P. Mukhopadhyay, and S. A. Rajasegarar, "Application of artificial intelligence in air pollution monitoring and forecasting: A systematic review," *Environmental Modelling & Software*, vol. 174, pp. 106–124, 2024.

[3] H. Abdel-Raouf and A. Youssef, "A systematic survey of air quality prediction based on deep learning," *Alexandria Engineering Journal*, vol. 67, no. 3, pp. 451–472, 2024.

[4] C. C. Zhang and L. Guo, "Machine learning approaches for outdoor air quality modelling: A systematic review," *Applied Sciences*, vol. 8, no. 12, pp. 2570–2588, Dec. 2018.

[5] K. V. Reddy, S. Mishra, and R. S. Rao, "A review on emerging artificial intelligence techniques for air pollution forecasting: Fundamentals, application and performance," *Journal of Cleaner Production*, vol. 320, pp. 128–145, 2021.

[6] A. Sharma and N. Patel, "Applications of machine learning and IoT for outdoor air pollution monitoring and prediction: A systematic literature review," *arXiv preprint arXiv:2401.01788*, 2024.

[7] Y. Wang, M. Li, and P. Kumar, "Systematic review of machine learning and deep learning techniques for spatiotemporal air quality prediction," *Atmosphere*, vol. 15, no. 11, pp. 1352–1369, 2024.

[8] J. Lee, K. H. Park, and T. Nguyen, "A review of machine learning for modeling air quality: Overlooked but important issues," *Atmospheric Research*, vol. 292, pp. 106–120, 2024.

[9] S. Li, A. Chen, and H. Sun, "Air quality forecasting with artificial intelligence techniques: A scientometric and content analysis," *Environmental Modelling & Software*, vol. 149, pp. 105–117, 2022.

[10] P. Das, A. Dutta, and V. Sharma, "Explainable AI insight: An orderly survey," in *Recent Trends in Computing*, LNNS vol. 748. Springer, 2023, pp. 123–140.

[11] N. Gupta and R. K. Singh, "A survey of explainable artificial intelligence for smart cities," *Electronics*, vol. 12, no. 4, pp. 1020–1035, 2023.

[12] M. I. Khan and J. P. Li, "Enhancing accuracy in urban air quality prediction: A comparative study of predictive algorithms for air pollutant concentrations," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 12, no. 2, pp. 78–89, 2024.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 12 , December - 2025

[13] K. Sundar and P. R. Kumar, “An intelligent IoT-cloud-based air pollution forecasting model using univariate time-series analysis,” *IEEE Access*, vol. 11, pp. 45678–45689, 2023.
[14] X. Liu, Z. Chen, and R. Wu, “High-resolution air quality prediction using low-cost sensors,” *arXiv preprint arXiv:2006.12092*, 2020.
[15] H. Luo, Y. Zhang, and C. Wu, “Federated learning in the sky: Aerial-ground air quality sensing framework with UAV swarms,” *arXiv preprint arXiv:2007.12004*, 2020.
[16] M. F. Hossain, T. Zaman, and M. A. Rahman, “A comprehensive survey on Internet of Things (IoT) based air pollution monitoring systems,” *Sustainable Computing: Informatics and Systems*, vol. 37, pp. 100–121, 2023.
[17] R. Dutta, K. K. Singh, and S. S. Shukla, “Artificial intelligence for sustainable air quality management: A review of techniques, challenges, and future directions,” *Environmental Science and Pollution Research*, vol. 31, no. 2, pp. 1450–1472, 2024.
[18] C. Li, X. Wang, and P. Zhang, “Review of deep learning applications in air quality prediction: Current trends and future perspectives,” *Ecological Informatics*, vol. 77, pp. 102–122, 2023.
[19] S. Banerjee, N. Mehta, and V. Kumar, “Air quality prediction using hybrid models: A systematic review of machine learning and statistical approaches,” *Journal of Environmental Management*, vol. 335, pp. 117–139, 2023.
[20] Y. K. Gaur, R. K. Mishra, and T. Gupta, “Explainable artificial intelligence for environmental monitoring: A systematic review,” *IEEE Access*, vol. 12, pp. 45690–45715, 2024.